拓扑感知的分层对等 Overlay 网格架构及其聚类算法研究

曹怀虎， 余镇危， 王银燕
（中国矿业大学北京校区机电与信息工程学院，北京，100083）

摘要：在考虑网格环境的异构性和逻辑拓扑时延性的基础上，提出了一种拓扑感知分层对等Overlay网格模型THP2POG架构，并对THP2POG聚类问题进行了形式化的描述，给出了一个基于预分组的聚类算法。实验表明该聚类算法可扩展性好，能有效地应用于网格环境中。
关键词：网格架构；拓扑感知；聚类算法
中图法分类号：TP393

Research on Clustering Algorithm in Topology-aware Hierarchical P2P Overlay Grid Architecture

CAO Huai-hu, YU Zhen-wei, WANG Yin-yan
(China University of Mining and Technology-Beijing, Beijing, 100083)

[Abstract] This paper propose THP2POG，a Grid Architecture which takes account of locality and heterogeneity of network hosts. To construct THP2POG with topology-aware，we design a clustering algorithm by take advantage of static and dynamic landmark algorithm. Simulation results show that the algorithm is scalable，simple and can be deployed effectively in Overlay Grid environment。
[Keywords] Overlay Grid Architecture； Topology-aware；Clustering algorithm

1. 引言

网格技术出现于 20 世纪 90 年代中期，当时是为了高级科学发现与工程研究而提出的分布式计算基础设施，时至今日，这种技术已经取得了相当大的进展，融合了从网络到人工智能的许多技术。网格是一种重要的技术趋势，所要解决的问题是，在动态的多制度的虚拟组织之间协调的资源共享与操作，这里的共享是指直接访问计算机、软件、数据和其它资源，而不单是指文件交换。

传统的网格都是直接基于物理网络或者专用网络来进行研究和构造的，需要网络底层硬件和协议的支持，由于实际的物理网络的异构、不确定、分布、自治、动态以及演化特性，使得网格计算变的过于复杂而不可行，并且不能保证其开放性和扩展性。

近年来出现的位于传统TCP/IP模型中的应用层P2P技术，是建立在IP层之上的应用层Overlay Network（Application-level overlay network），Overlay Network是一种构造网络的方法，它可在原有物理网络的基础上，通过构造一个虚拟网络，来支持原有网络没有或很难提供的功能，并能最大限度地保证与原有网络的兼容性[1]。因Overlay特殊的网络构造方式，使Overlay Network不需得到网络中所有组件的支持且无需改变已有的网络结构，即可为新型应用提供所需的服务。使网络资源更加易于控制和管理，同时也增强了网络的安全性能[2]。P2P技术已经拥有比较成熟的资源查找和路由策略，对于基于洪泛的P2P，如Gnutella等虽然支持完全分布式的查找策略，使得整个对等网络具有高鲁棒性，但洪泛的方式随着路由层数的深入，冗余消息数量将呈指数增长，极大的增加了网络的通信负担，可扩展性差。对于基于分布式哈希表的P2P，如Chord等则将查找对象转变为查找对象的位置信息，仅仅通过独立于对象位置的对象名，就能在可估上限的逻辑路由跳数内完成查询请求的路由，并定位到最近的存有对象副本的节点上，比较有效的解决了P2P可扩展性的问题，但是其维护开销要大于基于洪泛的P2P，并且不能支持模糊查询或更复杂的查询方式。上述经典P2P算法都是一种平面结构，系统中所有的节点都赋予相同的责任。而在评价各种P2P路由算法时，也主要以消息经过的逻辑路由跳数作为衡量一个算法性能的重要指标。然而现实网络中节点能力和会话时间具有极大的异构性，但是P2P中节点间的逻辑链路独立于实际的物理链路，所以

路由时仅仅考虑经过的节点数量是不够的，还必须考虑路由时的实际物理链路延迟。

在考虑节点的异质性和逻辑拓扑时延性的基础上，本文提出了一种拓扑感知分层对等Overlay网格架构THP2POG（Topology-aware Hierarchical P2P Overlay Grid），并对THP2POG聚类问题进行了形式化的描述，给出了一个基于预分组的聚类算法。

2. 拓扑感知的分层对等 Overlay 网格模型

2.1 THP2POG 模型的结构

基于 TCP/IP 的 Internet 取得巨大成功的主要原因在于其体系结构的开放与互连。通过采用层次化结构的参考模型，遵循端到端的原则,在很大程度上简化了网络协议的建模，并使得不同的网络系统可以方便的相互连接。

当前的 Internet 由于网络上设备数目庞大，网络管理复杂，引入了自治系统的概念，在拓扑上采取的是一种层次模型。自治系统内包含多个网络和非核心网关，通过内部的路由协议（例如 RIP，OSPF）完成路由，而核心网关间则通过边界路由协议（例如 BGP）交换路由信息，以保证整个 Internet 路由的一致性。Internet 的这种层次模型，使得各个自治系统的内部结构及路由算法对 Internet 的其它部分保持了透明性，既保证了整个 Internet 路由的一致性，又简化了数据报的处理，并极大的增加了 Internet 路由的灵活性。

而经典P2P算法中每个节点加入对等网络时都必须按照同一种分配算法分配逻辑空间地址（例如Gnutella的GUID、Chord中哈希后的节点关键字），按照同一种路由算法完成逻辑路由功能。

受Internet层次模型的启发，并反过来充分利用节点间的异构性，将Internet的层次模型引入到网格的构造中来，构造了一个拓扑感知的分层对等Overlay网格THP2POG（Topology-aware Hierarchical P2P Overlay Grid），其结构如图 1 所示，首先从功能模型上抽象为四层结构，其中：

（1） 基础网络层利用 TCP/IP 协议族在物理链路上提供基本的物理层传输、网络通信和 IP 路由等功能；

（2） 对等路由层按某种路由算法生成节点的路由表，为上面的数据管理层和应用层提供基于P2P的网格逻辑路由功能；

（3） 数据管理层则为对等系统内节点上的数据对象提供管理功能，例如数据存储、动态对象复制、负载平衡等；

（4） 应用层负责为用户提供多样化的应用服务；

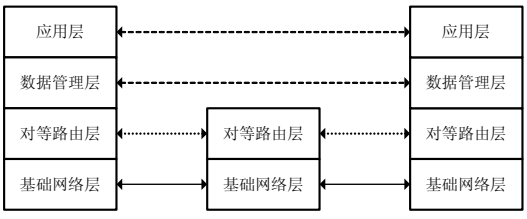


图 1 THP2POG 四层模型

其次，THP2POG将对等结点在延迟度量空间上分成若干个组，在结点加入网格时，先利用某种聚集算法寻找到自己所属的组，从而使得Overlay网格的逻辑拓扑尽量接近于实际的物理拓扑。之后每个分组内则选举出若干个节点作为超级节点，赋予其更多的功能，利用超级节点完成类似核心网关的功能，允许在不同的分组间采用不同的路由算法，从而提高Overlay网格的灵活性。图 2 给出了一个两层的THP2POG模型示意图。

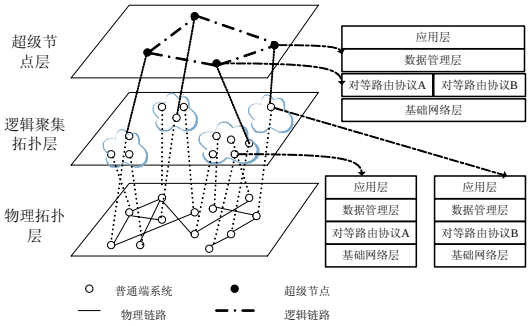


图 2 两层 THP2POG 分层模型

2.2 THP2POG 问题描述

按照上述定义，一个 K 分组的 THP2POG 可以表示为无向图 G=（V，E，W），其中 V 为对等结点的集合，E 为对等结点间逻辑链路的集合，W（u，v）（u，v∈V）表示逻辑链路的延迟，并满足：

（1） $V_m \cap V_n = \Phi$，$1 \leqslant m, n \leqslant K$；

（2） $\bigcup_{i=1}^{K} V_i = V$；

（3） $W(u,v) < W(u',v')$ iff $(u,v \in V_m) \wedge (u' \in V_p) \wedge (v' \in V_q) \wedge (p \neq q)$, $1 \leqslant m, p, q \leqslant K$；

3. 基于预分组的聚类算法

结构化 P2P 中为解决逻辑拓扑和物理拓扑不一致的问题，提出了三类算法：（1）邻近邻居选择路由算法在生成节点的路由表时，将路由表的指针优先指向物理网络邻近的节点以改善路由性能。（2）邻近路由算法则在每次路由选择下一路由跳时，采用启发式算法在路由表中选中一个在 ID 空间能有效地接近目标节点，并且时延开销又不太大的节点作为下一次路由跳。（3）而地理布局路由算法则在 Internet 上放置 M 个路标，按照对等节点与路标的相对距离分配相应的 ID。

这三种方法都是基于结构化 P2P 提出的，并且前两种方法有效性与节点的可选邻居数量直接相关，后一种算法则破坏了结构化 P2P 中节点在 ID 空间的均匀分布。

在 THP2POG 模型中，不论是基于无结构 P2P 还是基于结构化的 P2P，当节点加入 overlay 网格时，以延迟为度量空间寻找到自己所属的组，从而使得对等网格的逻辑拓扑尽量反映实际的物理拓扑，之后不同的 P2P 路由算法则可以充分利用每个分组中节点物理上邻近的特点。例如，同样采用结构化算法，对不同的分组采用不同的哈希函数分配节点的逻辑地址，则既可以解决逻辑拓扑和物理拓扑不一致的问题，又不会破坏节点在 ID 空间的均匀分布。

因此 THP2POG 模型中首先需要解决的问题是在节点加入系统时，采取何种聚类算法对节点进行聚集，以便将地理上邻近的节点分到同一个聚类中。

3.1 算法设计目标

作为一个应用于网格的聚集算法，首先必须要考虑到网格中节点都是端系统，并且数目庞大、动态性强的特点，所以先给出算法的设计目标：

（1） 实用性：算法应该容易实现，即不能有太多的假设条件；

（2） 可扩充性：算法必须是分布式算法，而不能假设由某个或某几个节点来收集全局的拓扑信息；

（3） 实 时 性：由于网格系统中节点的动态性，也就是加入/退出的不稳定性，聚集算法

不能太复杂，否则将引起系统的震荡；

3.2 拓扑信息的获取

为了获知节点之间在地理上的关系，一种方法是通过某个全局节点或第三方服务来收集全局的拓扑信息[1]。假定每个自治系统都有一个 SNMP 网管，通过该网管可以获得自治域内的静态拓扑结构。这种方法依赖网络中假设的全局节点，不能适用大规模的动态网络，可扩展性比较差[2]。则直接从 BGP 路由器上获取拓扑信息。但是由于 Internet 网络中端系统没有路由器的访问权限，所以直接从路由器上获取拓扑信息的方法对普通端系统根本无法适用。

另外还可考虑利用Traceroute工具，但Traceroute是通过每次发送一个TTL递加的数据报来完成探测目的，所以该方法是一种重量级方法，将会加重网络的负载。[3,4]使用Ping工具通过RTT（Round-trip time）测量来判断拓扑信息，缺点是判断粒度比较粗，但该方法的优点是轻量级、快速、实用，因此我们的聚类算法中也直接将延迟做为评价指标。

不过基于静态路标[3]的算法在节点数目多的时候，将使得充当静态路标的节点成为热点。[4]设计了一种动态路标算法来解决这个问题。然而这种通过路标探测的算法，路标的数量和在网络上的地理分布将直接关系到聚类的质量，而[4]在聚类初期将严重受制于其充当动态路标节点的数量，并且其聚类标准使得分组数量太大，很大程度上失去了分组的意义，后面的模拟实验也证明了这个问题。此外其组邻居表中邻居节点都是距离自己较近的其它组节点，容易出现网络分割的情况。

与本文的算法比较相似的是[5]，采取的也是先部分节点预分组再全局聚类的算法。但其通过计算初始集合中全部节点的距离矩阵来对初始集合进行分组，开销太大，并且固定分组数目的方法将无法适应大规模的网格计算。

3.3 聚类算法

我们的聚类算法主要分为两步：预分组和动态分组[6]。即在分组数目小于某阀值时以静态节点作为路标进行预分组，否则以分组超级节点的组邻居作为路标进行动态分组，并且都以随机游走（Random Walk）的方法生成组邻居表。

描述算法之前，首先给出分布式编码定义：

定义 1：令L表示为m个路标的集合，$L=\{L_0, L_1, L_2, ...L_{m-1}\}$，$L_a$是节点a在L上关于$L_i$（$0 \leqslant i \leqslant m-1$）的一个排序，且满足：

$i1<i2$ iff $(RTT(a,L_{i1})<RTT(a,L_{i2})) \vee (((RTT(a,L_{i1})=RTT(a,L_{i2})) \wedge (L_{i1}<L_{i2}))$

先选择若干个比较知名的网站作为静态路标。每个分组中的超级结点都保存两个表：邻居表GNT（Group Neighbor Table）和成员表GMT（Group Member Table）。系统预分组数目阀值为 ε，GNT最大允许数目为M_{GNT}，邻居搜索终止步数为T。

当新节点 a 加入网格系统，假设其已知系统中某个结点 z（这可以通过采用集合点或洪泛等多种办法实现）：

（1） 通过 z 找到其所属组的超级结点 SZ；

（2） 如果 SZ 的 GNT 数目大于 ε，转（5）；

（3） 以静态路标作为路标集合 L，分别对 a 和 SZ 及其 GNT 中节点计算其各自的分布式编码；

（4） 如果$L_a=L_{SZ}$或者$L_a=L_{GNT(i)}$，a属于SZ或者GNT（i）所在分组，将a加入SZ或者GNT（i）的GMT中，算法结束。否则在SZ及其GNT中探测距离自己最近的超级节点NS，如果NS≠SZ，将SZ置为NS，重复（3）（4）。否则创建一个新组，a作为新组的超级节点，邻居搜索步数置为 0，转（7）；

（5） 以 SZ 的 GNT 作为路标集合 L，分别对 a 和 SZ 计算其各自的分布式编码；

（6） 如果$L_a=L_{SZ}$，a属于SZ所在分组，将a加入SZ的GMT中，算法结束。否则在SZ及其GNT中探测距离

自己最近的超级节点NS，如果NS≠SZ,将SZ置为NS，重复（5）（6）。否则创建一个新组，a作为新组的超级节点，邻居搜索步数置为 0；

（7） 邻居搜索步数加 1。若SZ不在a的GNT中且SZ的GNT数目小于M_{GNT}，a和SZ分别将对方加入自己的GNT中；

（8） 从SZ的GNT中随机取一超级节点RS，若a的GNT数目小于M_{GNT}并且邻居搜索步数不大于T，将SZ置为RS重复（7）（8），否则算法结束；

上述算法描述中，（3）（4）为预分组算法，（5）（6）为动态分组算法，（7）（8）是超级节点邻居表生成算法。可以看出，不管是预分组还是动态分组，新节点总是向距离自己更近的分组超级节点靠近直到找到所属分组或者创建新组为止，所以该算法是收敛的。

4. 模拟实验

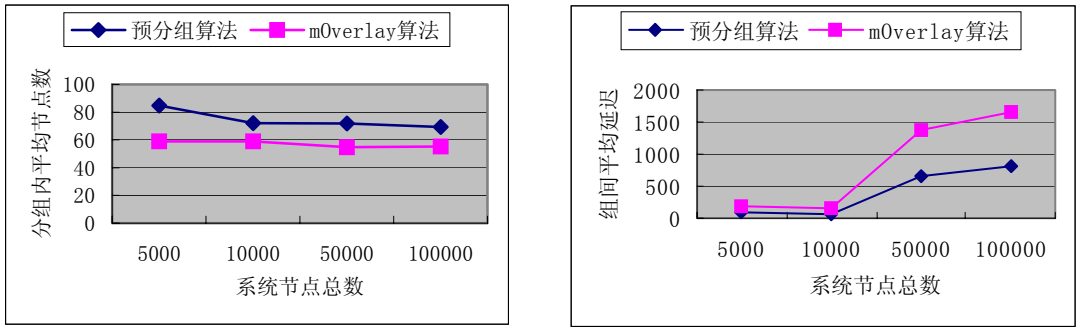
为了比较不同算法的性能，我们在Brite的Java开源版本的基础上[7]，实现了基于预分组的聚类算法和mOverlay [4]中的聚类算法，其中节点间延迟用拓扑图中Dijkstra算法计算出来的节点间最短距离来代替。

拓扑图分别采用WAXMAN，BA以及GLP模型在Brite中以路由器级生成[8]，图中每个点和 [4]实验一样代表 50 个端系统，并在生成的拓扑图中手动选取了若干个度数较高的节点来模拟静态路标节点[9]。

限于篇幅，下面仅给出在 GPL 拓扑生成模型下四组实验结果。系统总节点数分别为 5000、10000、50000 和 100000 个节点。GPL 模型中 p 设置为 0.5，beta 设置为 0.64，M 设置为 3，带宽分布为重尾分布（Heavy Tailed）[10]。

两种聚类算法中组邻居表M_{GNT}都为 10，邻居搜索终止步数T都为 20。预分组算法中静态路标个数为 12，预分组数目阀值 ε 为 10。

图 3（a）中给出的是聚类内平均节点数目比较结果，从图中可以看出，预分组算法分组内平均节点数目比 mOverlay[4]算法提高了近 32%。而图 3（b）的实验结果中，预分组算法的组间平均延迟随着节点总数的增加将明显优于 mOverlay[4]算法。



（a） （b）

图 3 GPL 拓扑中聚类算法实验比较结果

5. 结论

节点的异质性和逻辑拓扑的时延性是网格所面临的一个现实问题，而应用的多样性也促使研究人员从协议框架的角度来认识网格技术。

本文在考虑上述问题的基础上，提出了一种拓扑感知分层对等Overlay网格模型THP2POG架构，并对THP2POG聚类问题进行了形式化的描述，并且为了使THP2POG模型中节点的逻辑拓扑能尽量反映物理拓扑，给出了一个基于预分组的聚类算法，实验的结果表明该聚类算法可扩展性好，能有效的应用于网格环境中。

参考文献

1．黄道颖，黄建华，庄雷，等.基于主动网络的分布式 P2P 网络模型[J].软件学报，2004，15(7)：1081-1089

2．]Balachangder Krishnamurthy，Jia Wang. On network-aware clustering of web clients[A]. In Proc. of ACM SIGCOMM2000[C]，2000

3．S.Ratnasamy，M.Handley，R.Karp，et al. Topologically-aware overlay constructions and server selections[A]. In Proc. of INFOCOMM2002[C]，2002

4．Xin Yan Zhang，Qian Zhang， Zhengsheng Zhang，et al. A construction of Locality-Aware Overlay Network: mOverlay and Its Performance[A].IEEE Journal on Selected Areas in Communications[C]，2004，22(1)：18-28

5.徐非，杨广文，鞠大鹏.基于 Peer-to-Peer 的分布式存储系统设计[J].软件学报，2004，15(2)：268-277

6. Czajkowski, K., Foster, I., Karonis, N., Carl Kesselman, C., Martin, S., Smith, W., and Tuecke, S. A Resource Management Architecture for Metacomputing Systems .Proc.Workshop on Job Scheduling Strategies for Parallel Processing, pp 62-82, Springer-Verlag,ISBN 3-540-64825-9, 1998.

7. Smith, P., Simpson S., Hutchison, D.Peer-to-Peer Networking for Discovering Programmable Resources, Proc. 4th Intl. Workshop on Networked Group Communication (NGC 2002), 23rd - 25th October 2002, Boston, USA.

8.. Pallickara, S., Fox, G.. NaradaBrokering: A Distributed Middleware Framework and Architecture for Enabling Durable Peer-to-Peer Grids, Proc. IFIP/ACM/USENIX Middleware 03, Rio de Janeiro, Brazil, April 2003.

9. M.L. Fisher. The lagrangean relaxation method for solving integer programming problems. Management Science, 27(1):1-18, 1981.

10. Frank K. Hwang, Dana S. Richards, and Pawel Winter. The Steiner Tree Problem. North-Holland, 1992.

研究背景

基金项目：高等学校博士学科点专项科研基金(20030290003) 项目名称：Overlay 网络服务体系的研究。

曹怀虎(1977－),男，博士研究生，研究方向：新一代网络体系结构，网格技术，Overlay Network。chhu@cumtb.edu.cn，北京海淀区学院路丁 11 号中国矿业大学机电学院博 03-2，邮编 100083

余镇危（1942－），男，博士生导师，主要从事网络体系结构，计算机系统结构，Overlay Network，ATM 核心技术方面的研究。